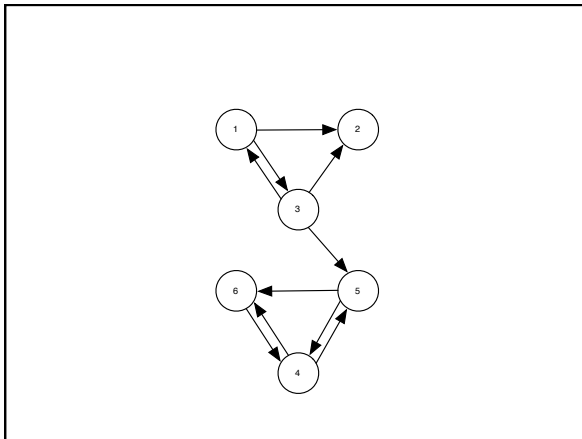


## Informationsgenerierung und -Speicherung




---

---

---

---

---

---

---

---

**HITS Beispiel**

6 Dokumente:

Dokument #	Hat ein Link zum Dokument(e) #
1	2,3
2	-
3	1,2,5
4	5,6
5	4,6
6	4

---

---

---

---

---

---

---

---

### Nichteindeutigkeit der HITS Lösung

- Auch wenn HITS immer konvergiert, gibt es ein Problem mit der Eindeutigkeit der Lösung der Hub und Authorities.
- Man kann zeigen, dass für verschiedene Anfangswerte die Power Method verschiedene Lösungen für Authority und Hub zurückgeben können.

---

---

---

---

---

---

---

---

### Das Community Problem

- HITS findet die **Dominanteigenvektoren [principal, dominant eigenvectors]**. Diese stellen die Regionen des Graphs dar, die am dichtesten verbunden sind **für eine spezifische Abfrage**.
- Für gewisse Fälle wäre es auch interessant, verschiedene, getrennte dicht verbundene Regionen zu finden, basierend auf einem spezifischen Base Set.
- Solche Regionen könnten potentiell beide relevant sein für die Abfrage, aber sehr deutlich geteilt sein im Graph.

---

---

---

---

---

---

---

---

### Community problem: Beispiel

Mehrdeutige Abfragen, z.B., "jaguar" entweder eine Katze oder ein Auto.

Polarisierende Themen, wo Gruppen involviert sind, die wenig zusammen zu tun haben, z.B. "Abtreibung".

---

---

---

---

---

---

---

---

### HITS Communities

- Für jedes Beispiel sind die relevante Seiten in verschiedene Clusters gruppiert, die Communities genannt werden
- Diese kleinere Clusters können mittels Berechnung von nichtdominante Eigenvektoren gefunden werden.
- Diese Eigenvektoren wiederum können ähnlich wie mit der Power Iteration Method gefunden werden, wie z.B. die orthogonale Iteration oder QR-Iteration.

---

---

---

---

---

---

---

---

### Beziehung mit Ko-Zitierung und Bibliographische Kopplung

- Authority Seiten und Hubseiten funktionieren ähnlich wie das Zitieren von wissenschaftliche Publikationen. Eine Authorityseite ist wie ein einflussreiche Publikation, die in vielen nachfolgenden Publikationen zitiert wird. Eine Hubseite ist wie ein Übersichtsartikel, der viele andere Publikationen zitiert.
- Ko-zitieren in der bibliographischen Welt ist genauso definiert wie die Authority oder Hub Matrix:

$$C_{ij} = \sum_{k=1}^n L_{ki} L_{kj} = (\mathbf{L}^T \mathbf{L})_{ij}.$$

- Damit ist die Authority Matrix ( $\mathbf{L}^T \mathbf{L}$ ) von HITS die Ko-Zitierungsmatrix C im Webkontext

---

---

---

---

---

---

---

---

### Beziehung mit Bibliographische Kopplung (1963)

- Die bibliographische Kopplung von 2 Seiten  $i$  und  $j$ , sind mit  $B_{ij}$  beschrieben und berechnet als:

$$B_{ij} = \sum_{k=1}^n L_{ik} L_{jk} = (\mathbf{L} \mathbf{L}^T)_{ij}, \quad (41)$$

- Damit ist die Hub Matrix ( $\mathbf{L} \mathbf{L}^T$ ) von HITS gleich wie die bibliographische Kopplung B, im Webkontext platziert.

---

---

---

---

---

---

---

---

### Stärke und schwäche von HITS

- Die Hauptstärke von HITS ist die Fähigkeit, Seiten zu ordnen in Abhängigkeit zu einem bestimmten Thema, damit kann HITS drastisch die Relevanz erhöhen. Das kann mit Standard IR Methoden kombiniert werden, um die Performanz einer Methode zu erhöhen.
- Eine weitere Stärke ist dass HITS eigentlich 2 Klassifikationen gleichzeitig liefert. Die Autoritäten sind gut für Vertiefungen in einem Thema, während die Hubs sind für breite Suchen gut.
- Eine weitere Stärke ist, dass HITS immer mit relativ kleine Matrizen arbeitet im Vergleich mit der Größe vom Web.
- Die Hauptschwäche ist, dass HITS muss für jede Abfrage evaluiert werden, damit wird die Evaluationszeit eine echte Hürde. (Wie kann dieses Problem gelöst werden?)
- Eine weitere Schwäche ist die Empfindlichkeit zu Spam. Eine Person kann einfach auf einer Seite Links addieren und somit die Hub-Score beeinflussen. Da aber der Hub und Authority verbunden sind, wird auch der Authority score auch sich vergrößern, wenn der Hub score sich vergrößert (es gibt aber Lösungen für das).
- Eine weitere Schwäche ist das „Topic Drift“: wenn der Base Set aus dem Root Set erweitert wird, kann es sein, dass eine sehr autoritative Seite, jedoch nicht voll zur Abfrage verbundene Seite gelinkt wird. Diese Seite kann dann das Resultat in einer irrelevanten Richtung verschieben.

---

---

---

---

---

---

---

---